Summary of ARC’s CDDIS Metadata Findings

March 21, 2017

This report outlines the ARC team’s metadata findings for the CDDIS DAAC. The findings reported below are present in the majority of the collection level metadata records. Detailed reports for individual collection and granule level metadata records are provided separately. The detailed reports will be used for tracking the DAAC’s progress towards working off the reported findings and metrics will be generated from the detailed reports. An explanation of the color coding found in the detailed reports is included in this report along with preliminary metrics.

1. Collection Level Findings

i. If possible, leverage the EOSDIS Earthdata login mechanism. Earthdata login provides a centralized mechanism for user registration across EOSDIS system components. It is ESDIS policy that all DAAC and ESDIS applications where data from data products are retrieved by humans or machines use Earthdata Login. More information on the Earthdata login, including information on how to integrate Earthdata login into a system, can be found at the EOSDIS User Registration System wiki page (https://wiki.earthdata.nasa.gov/display/URSFOUR/EOSDIS+User+Registration+System)

ii. CDDIS currently provides data access via FTP. Data access is preferably provided via HTTP instead of FTP. ARC recommends the migration of data access to HTTP whenever possible in order to encourage DAAC alignment with the NASA requirement. ARC is aware of CDDIS’s large FTP user community and acknowledges the effort required in order to retire FTP at CDDIS.

iii. Whenever feasible, the Online Access URL in the metadata should point as directly as possible to the described data. For example, some of the GNSS collection level metadata includes an online access URL that points to the product level ftp folder. As a new user, it can be difficult to determine which sub-folder is the correct one for the selected metadata record. Pointing more directly to the appropriate folder eliminates confusion for the user and eases data accessibility.

iv. Whenever possible, a link to the dataset landing page should be included for each collection level record. Ideally this link will leverage the DOI URL. More information on DOI landing pages and the information required for a dataset landing page can be found here: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/DOI+Landing+Page

v. Please include links to relevant documentation for each collection level metadata record. Relevant documentation could include direct links to the CDDIS data summary reports (https://cddis.nasa.gov/Data_and_Derived_Products/Reports.html) and also the CDDIS data citation page.

vi. Please include links to relevant data discovery tools (https://cddis.nasa.gov/Data_and_Derived_Products/Data_Discovery.html) for each collection level metadata record.

vii. All collection level metadata records should include a unique and descriptive long name or title for each dataset. The long name in the collection level metadata should match the dataset long name found on the DOI landing page (https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/DOI+Landing+Page). Descriptive and unique dataset titles are important for data discovery.

viii. Whenever possible, the long name or title should be included in the first sentence of the abstract. Providing the title in the first sentence of the abstract provides clarity for a user who discovers the dataset via the CMR API or has downloaded the metadata directly.

ix. Several collection level metadata records are missing science keywords. At least one science keyword is required for all collection level metadata records. Science keywords should reflect scientific parameters provided in the data and may also include more broad keywords used to describe the data set. ARC has provided some suggested science keywords for most collection level records in the detailed reports. Please add as many GCMD science keywords as necessary to both describe the dataset and to also aid users in discovering the data. GCMD keywords can be found here - http://gcmdservices.gsfc.nasa.gov/static/kms/sciencekeywords/sciencekeywords.csv?ed_wiki_keywords_page. Please let us know if more guidance is needed. We will be happy to assist.

x. The evaluation of science keywords for CDDIS data sets revealed the potential for several new keywords which could be added to the GCMD KMS in order to better support the type of data CDDIS provides. Keyword suggestions are summarised in the following bullets. Integrating new keywords into a KMS hierarchy (if desired) will require collaboration between CDDIS and GCMD. ARC can provide guidance and support for this collaborative process as necessary.

- "Laser Lunar Ranging" or a similar keyword could be added for LLR datasets.
- There is currently an "ORBITAL CHARACTERISTICS" keyword. Perhaps this keyword could be expanded to specify the particular orbital characteristics, such as satellite altitude, orbital inclination, etc. which are relevant to the SLR data sets.
- "Zenith Path Delay" could be added as a keyword to better describe the GNSS troposphere and ionosphere products.
- There is a "ORBITAL POSITION" keyword yet there are no keywords specific to Orbital Position Predictions/ Orbital Projections. These specific keywords could be added to accommodate the SLR Satellite Prediction data sets.
- A keyword related to measuring time could be added to GCMD for the clock data sets. Currently there is a "DATA SYNCHRONIZATION TIME" keyword but this is meant for aircraft data.
- "Celestial Reference Frame" and/or "Quasars" could be added to accommodate VLBI data sets. In addition, "Very Long Baseline Interferometry" could be added as an instrument keyword for VLBI data sets, although "RADIO TELESCOPES" is an already existing alternative.

xi. Data format information should be included in all collection level metadata records. Data format information is important in helping a user determine whether he or she can use the discovered data.

xii. 'Collection state' is now a required element in the Unified Metadata Model (UMM). More specific guidance on collection state information for each collection is provided in the detailed report. Additional information on 'collection state' within the UMM can be found here - https://wiki.earthdata.nasa.gov/display/CMR/CMR+Documents

xiii. To achieve naming consistency, it is recommended that the Archive Center name be compliant with GCMD 'Data Centers' vocabulary. Currently "NASA/GSFC/SSED/CDDIS" is the only representation of CDDIS in the GCMD Data Centers list: http://gcmdservices.gsfc.nasa.gov/static/kms/providers/providers.csv. All metadata records from the same DAAC should have the same archive center name for consistency. Should CDDIS wish to use only "CDDIS" as the archive center name, CDDIS should contact GCMD to coordinate changes to the keyword list.

2. Granule Level Findings:

i. The ‘Measured Parameter’ field currently lists the data product name. According to the ECHO 10 Data Partner User Guide, the measured parameters field is meant to contain the names of the geophysical parameters expressed in the data.
Therefore, CDDIS can consider improving this field by replacing the product name with the observed geophysical parameters in the data, or removing this element from the granule level metadata.

ii. In the granule level metadata, there is an element called ‘Size MB Data Granule.’ The name of the element limits the listing of the granule file size to megabytes only. It appears that the majority of granule metadata records checked included the file size in kilobytes, rather than in megabytes.

iii. Please include links to relevant resources for each granule level metadata record. Relevant resource links could include links to the CDDIS data citation page and links to the appropriate data discovery tools (https://cddis.nasa.gov/Data_and_Derived_Products/Data_Discovery.html).

3. Explanation of Color Coding Provided in Detailed Reports

Color	**Definition**
Cyan	Required field based on UMM-C
Light Purple	An optional primary element with required sub-elements based on UMM-C
Purple	A sub-element which is only required if any information is provided in the scope of the primary element based on UMM-C
White	Completely optional field
Red	Correcting these issues should be of the highest priority
Yellow	Correcting these errors are strongly recommended but are not required
Blue	Minor error/ inconsistency; points out

🟦	features noticed by the ARC Team which may help improve the robustness of the metadata but are not required to be addressed
*	Any field with an asterisk is controlled by GCMD vocabulary

4. Metrics

38 Collection level records checked

Collection Level	# Red fields	# Yellow fields	# Blue fields	Total # fields checked
	481	257	144	2,031
	23.7%	12.7%	7.1%	

28 Granule level records checked

Granule Level	# Red fields	# Yellow fields	# Blue fields	Total # fields checked
	36	92	90	1,545
	2.3%	6.0%	5.8%	

66 Total records checked (collection + granule)

Cumulative	# Red fields	# Yellow fields	# Blue fields	Total # fields checked
	517	349	234	3,576
	14.5%	9.8%	6.5%	